Factors for Constructing Control Limits in the `rQCC` package

Chanseok Park* and Min Wang†

December 2022

Abstract

In this note, we provide several mathematical formulas of the factors which are used for constructing the control limits. These factors can be easily obtained by using the `factors.cc` function in the robust quality control chart (`rQCC`) R package.

1 Factors for computing control chart lines

In this section, we provide a brief summary of mathematical relations of factors for computing the control chart *lines*. For more details, see Supplement A of ASTM (STP 15-D) [1] and Supplement B of ASTM (STP 15-C) [2].

The mathematical relations for factors (c_2, c_4, d_2, d_3) are based on sampling randomly from a normal distribution. These are given by

$$c_2(n) = \sqrt{\frac{2}{n}} \cdot \frac{\Gamma(n/2)}{\Gamma(n/2-1/2)},$$

$$c_4(n) = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(n/2-1/2)},$$

$$d_2(n) = 2\int_0^\infty \Big\{1 - \big[\Phi(z)\big]^n - \big[1-\Phi(z)\big]^n\Big\} dz,$$

and

$$d_3(n) = \sqrt{E(R^2) - d_2(n)^2},$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution and $E(R^k)$ is given in (2) of Appendix B. All the detailed derivations of $c_4(n)$ are provided in Appendix A and those of $d_2(n)$ and $d_3(n)$ are given in Appendix B. Note that $c_2(n)$ has been used in ASTM (STP 15-C) [2] and it is replaced by $c_4(n)$ in ASTM (STP 15-D) [1]. Thus, $c_2(n)$ is rarely used after the year of 1976.

*Applied Statistics Laboratory, Department of Industrial Engineering, Pusan National University, Busan 46241, Korea. His work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C1091319).

†Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX 78249, USA.

2 Factors for computing control limits

The factors below are used for constructing a variety of control charts with the choice of $g \cdot \sigma$ *limits*. For more details, see Supplement A of ASTM (STP 15-D) [1] and Supplement B of ASTM (STP 15-C) [2]. Note that the American Standard uses $3 \cdot \sigma$ limits with 0.27% false alarm rate, while the British Standard uses $3.09 \cdot \sigma$ limits with 0.20% false alarm rate.

- For averages:

$$\begin{aligned} A(n) &= \frac{g}{\sqrt{n}}, \\ A_1(n) &= \frac{g}{c_2(n)\sqrt{n}} = \frac{A(n)}{c_2(n)}, \\ A_2(n) &= \frac{g}{d_2(n)\sqrt{n}} = \frac{A(n)}{d_2(n)}, \\ A_3(n) &= \frac{g}{c_4(n)\sqrt{n}} = \frac{A(n)}{c_4(n)}. \end{aligned}$$

  Note that $A_1(n)$ in ASTM (STP 15-C) [2] was replaced by $A_3(n)$ in ASTM (STP 15-D) [1] in the year of 1976. Since then, $A_1(n)$ is rarely used.

- For standard deviations:

$$\begin{aligned} B_1(n) &= \max\left\{ c_2(n) - g \cdot \sqrt{\frac{n-1}{n} - c_2(n)^2},\ 0 \right\}, \\ B_2(n) &= c_2(n) + g \cdot \sqrt{\frac{n-1}{n} - c_2(n)^2}, \\ B_3(n) &= \max\left\{ 1 - \frac{g}{c_4(n)} \cdot \sqrt{1 - c_4(n)^2},\ 0 \right\}, \\ B_4(n) &= 1 + \frac{g}{c_4(n)} \cdot \sqrt{1 - c_4(n)^2}, \\ B_5(n) &= \max\left\{ c_4(n) - g \cdot \sqrt{1 - c_4(n)^2},\ 0 \right\} = c_4(n) \cdot B_3(n), \\ B_6(n) &= c_4(n) + g \cdot \sqrt{1 - c_4(n)^2} = c_4(n) \cdot B_4(n). \end{aligned}$$

  Note that $B_1(n)$ and $B_2(n)$ in ASTM (STP 15-C) [2] are replaced by $B_5(n)$ and $B_6(n)$, respectively, in ASTM (STP 15-D) [1].

  In ASTM (STP 15-C), however, $B_3(n)$ and $B_4(n)$ are defined as

$$\begin{aligned} B_3(n) &= \max\left\{ 1 - \frac{g}{c_2(n)} \cdot \sqrt{\frac{n-1}{n} - c_2(n)^2},\ 0 \right\}, \\ B_4(n) &= 1 + \frac{g}{c_2(n)} \cdot \sqrt{\frac{n-1}{n} - c_2(n)^2}, \end{aligned}$$

  which are easily obtained by $B_1(n)/c_2(n)$ and $B_2(n)/c_2(n)$ in ASTM (STP 15-C), respectively. Thus, we calculate $B_3(n)$ and $B_4(n)$ based only on ASTM (STP 15-D) [1] instead of ASTM (STP 15-C).

- For ranges:

$$
\begin{aligned}
D_1(n) &= \max\left\{d_2(n) - g \cdot d_3(n),\ 0\right\},\\
D_2(n) &= d_2(n) + g \cdot d_3(n),\\
D_3(n) &= \max\left\{1 - g \cdot \frac{d_3(n)}{d_2(n)},\ 0\right\} = \frac{D_1(n)}{d_2(n)},\\
D_4(n) &= 1 + g \cdot \frac{d_3(n)}{d_2(n)} = \frac{D_2(n)}{d_2(n)}.
\end{aligned}
$$

- For individuals:

$$
\begin{aligned}
E_1(n) &= \frac{g}{c_2(n)},\\
E_2(n) &= \frac{g}{d_2(n)},\\
E_3(n) &= \frac{g}{c_4(n)}.
\end{aligned}
$$

Note that $E_1(n)$ in ASTM (STP 15-C) [2] is replaced by $E_3(n)$ in ASTM (STP 15-D) [1].

Appendices

A The bias correction factor for the sample standard deviation

It is well known that

$$E(S^2) = \sigma^2,$$

where $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, $X_i \sim N(\mu, \sigma^2)$, and $\bar{X} = \sum_{i=1}^{n} X_i/n$. It deserves mentioning that $E(S) \neq \sigma$.

Using the fact that $Y = (n-1)S^2/\sigma^2$ has the chi-square distribution with $n-1$ degrees of freedom which is equivalent to the gamma distribution with $\alpha = (n-1)/2$ (shape) and $\theta = 2$ (scale), we obtain the unbiased estimator of σ. Now, it is well known that

$$E\big[Y^c\big] = \frac{\Gamma(\alpha + c)\theta^c}{\Gamma(\alpha)},$$

when Y has the gamma distribution with shape α and scale θ. Clearly, for $c = 1/2$, we have

$$E\big[\sqrt{Y}\big] = \frac{\Gamma(\alpha + 1/2)\sqrt{\theta}}{\Gamma(\alpha)}.$$

Then we obtain

$$E\big[\sqrt{(n-1)S^2/\sigma^2}\big] = \frac{\Gamma(n/2)\sqrt{2}}{\Gamma(n/2 - 1/2)}.$$

This implies that

$$E(S) = c_4(n)\,\sigma,$$

where

$$c_4(n) = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma(n/2 - 1/2)}.$$

Thus, the estimator $S/c_4(n)$ is unbiased for σ.

B The bias correction factors for the range

We can also estimate σ using the range, $R = X_{(n)} - X_{(1)}$, where $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ are the order statistics of a random sample of size n from $N(\mu, \sigma)$. It is known that $R = X_{(n)} - X_{(1)}$ by itself is *not* unbiased for σ. In this section, we provide the bias correction factor for the range to estimate σ so that $E(R/d_2(n)) = \sigma$. We also provide the bias correction factor defined by $\mathrm{Var}(R/d_3(n)) = \sigma^2$. First, we provide the following theorems and lemmas which are needed to obtain $d_2(n)$ and $d_3(n)$.

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be a random sample with continuous cdf $F(x)$ and pdf $f(x)$. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of a random sample. Then the joint pdf of $U = X_{(i)}$*

and $V = X_{(j)}$ *for* $1 \le i < j \le n$ *is given by*

$$f_{(i,j)}(u,v) = \frac{n!}{(i-1)! \times 1! \times (j-i-1)! \times 1! \times (n-j)!} \times \\ \Big[F(u)\Big]^{i-1} f(u) \Big[F(v) - F(u)\Big]^{j-i-1} f(v) \Big[1 - F(v)\Big]^{n-j}$$

for $-\infty < u < v < \infty$.

Proof. For more details, refer to Theorem 5.4.6 in Casella and Berger [3]. □

Let $Z_1, Z_2, \ldots, Z_n$ be a random sample from a standard normal distribution with the pdf $\phi(z)$ and the cdf $\Phi(z)$. For notational convenience, we denote $U = Z_{(1)}$ and $V = Z_{(n)}$. Using Theorem 1, we have the joint pdf of U and V

$$f_{(1,n)}(u,v) = n(n-1)\,\phi(u)\,\phi(v)\,\big[\Phi(v) - \Phi(u)\big]^{n-2}.$$

The goal is to derive the distribution of the range of the sample, $V - U = Z_{(n)} - Z_{(1)}$. Next we consider the new random variables given by $Y_1 = U$ and $Y_2 = V - U$. Notice that the random variable Y_2 is the *range*. The inverse transforms are easily obtained by $u = y_1$ and $v = y_1 + y_2$. Then the joint pdf of Y_1 and Y_2, denoted by $g(y_1, y_2)$, is given by

$$g(y_1, y_2) = n(n-1)\,\phi(y_1)\,\phi(y_1 + y_2)\,\Big[\Phi(y_1 + y_2) - \Phi(y_1)\Big]^{n-2} |J|,$$

where $-\infty < y_1 < \infty$, $y_2 > 0$ and

$$J = \det \begin{pmatrix} \dfrac{\partial u}{\partial y_1} & \dfrac{\partial u}{\partial y_2} \\ \dfrac{\partial v}{\partial y_1} & \dfrac{\partial v}{\partial y_2} \end{pmatrix} = 1.$$

Thus, we have

$$\begin{aligned} g_2(y_2) &= \int_{-\infty}^{\infty} g(y_1, y_2)\,dy_1 \\ &= n(n-1) \int_{-\infty}^{\infty} \phi(y_1)\phi(y_1 + y_2)\Big[\Phi(y_1 + y_2) - \Phi(y_1)\Big]^{n-2} dy_1. \end{aligned} \tag{1}$$

Note that the cdf of Y_2 can be easily obtained by

$$G_2(y_2) = n \int_{-\infty}^{\infty} \phi(y_1)\Big[\Phi(y_1 + y_2) - \Phi(y_1)\Big]^{n-1} dy_1.$$

Next we consider the k-th moment of the range which was provided by Harter [4]. We provide a detailed derivation here. Using the pdf of the range in (1), we can obtain the k-th moment of the range, $Y_2 = Z_{(n)} - Z_{(1)}$, by calculating the expectation as follows:

$$\begin{aligned} E(Y_2^k) &= \int_0^\infty y_2^k\, g_2(y_2)\, dy_2 \\ &= n(n-1)\int_0^\infty y_2^k \int_{-\infty}^\infty \phi(y_1)\phi(y_1+y_2)\Big[\Phi(y_1+y_2) - \Phi(y_1)\Big]^{n-2} dy_1\, dy_2 \\ &= n(n-1)\int_{-\infty}^\infty \left\{ \int_0^\infty y_2^k \Big[\Phi(y_1+y_2) - \Phi(y_1)\Big]^{n-2} \phi(y_1+y_2) dy_2 \right\} \phi(y_1)\, dy_1. \end{aligned}$$

For notational convenience, we replace (y_1, y_2) with (x, r). Then we have

$$E(R^k) = n(n-1)\int_{-\infty}^\infty \left\{ \int_0^\infty r^k \Big[\Phi(x+r) - \Phi(x)\Big]^{n-2} \phi(x+r) dr \right\} \phi(x)\, dx, \tag{2}$$

where $R = Z_{(n)} - Z_{(1)}$.

Clearly, the expression for the k-th moment of the range requires the evaluation of a complicated double integral. Fortunately, for the case where $k = 1$ which is the expectation, we can derive an alternative formula involving only a *single* integral. The derivation of this formula will require the application of three different lemmas which we state and prove below.

Lemma 2. *Let X be a continuous random variable with cdf $F(x)$. If $E(|X|^k)$ exists, then we have*

$$\textit{(i)}\ \lim_{x\to\infty} x^k\big\{1 - F(x)\big\} = 0 \quad \textit{and} \quad \textit{(ii)}\ \lim_{x\to-\infty} |x|^k F(x) = 0.$$

Proof. (i) For $x > 0$, we have

$$0 \le x^k\big\{1 - F(x)\big\} = x^k \int_x^\infty dF(t) = \int_x^\infty x^k dF(t) \le \int_x^\infty t^k dF(t).$$

Now, if we can show that the last term $\int_x^\infty t^k dF(t) \to 0$ in the limit as $x \to \infty$, then this will complete the proof because we just showed that $0 \le x^k\big\{1 - F(x)\big\} \le \int_x^\infty t^k dF(t)$ for $x > 0$. To prove that $\int_x^\infty t^k dF(t) \to 0$ in the limit as $x \to \infty$, we observe that

$$\int_x^\infty t^k dF(t) = \int_{-\infty}^\infty |t|^k dF(t) - \int_{-\infty}^x |t|^k dF(t) = E(|X|^k) - \int_{-\infty}^x |t|^k dF(t).$$

Since $E(|X|^k)$ exists and $\lim_{x\to\infty} \int_{-\infty}^x |t|^k dF(t) = E(|X|^k)$, we have

$$\int_x^\infty t^k dF(t) = E(|X|^k) - \int_{-\infty}^x |t|^k dF(t) \to 0$$

in the limit as $x \to \infty$.

(ii) For $x < 0$, we have

$$0 \le |x|^k F(x) = |x|^k \int_{-\infty}^x dF(t) = \int_{-\infty}^x |x|^k dF(t) \le \int_{-\infty}^x |t|^k dF(t).$$

Now, if we can show that the last term $\int_{-\infty}^{x} |t|^k dF(t) \to 0$ in the limit as $x \to -\infty$, then this will complete the proof because we just showed that $0 \le |x|^k F(x) \le \int_{-\infty}^{x} |t|^k dF(t)$ for $x < 0$. To prove that $\int_{-\infty}^{x} |t|^k dF(t) \to 0$ in the limit as $x \to -\infty$, we have

$$\int_{-\infty}^{x} |t|^k dF(t) = \int_{-\infty}^{\infty} |t|^k dF(t) - \int_{x}^{\infty} |t|^k dF(t) = E(|X|^k) - \int_{x}^{\infty} |t|^k dF(t).$$

Since $E(|X|^k)$ exists and $\lim_{x\to-\infty} \int_{x}^{\infty} |t|^k dF(t) = E(|X|^k)$, we have

$$\int_{-\infty}^{x} |t|^k dF(t) = E(|X|^k) - \int_{x}^{\infty} |t|^k dF(t) \to 0$$

in the limit as $x \to -\infty$. □

Lemma 3. *Let X be a continuous random variable with cdf $F(x)$. Then we have*

$$E(X) = \int_{0}^{\infty} \big[1 - F(x) - F(-x)\big]\,dx.$$

Proof. We have

$$E(X) = \int_{-\infty}^{0} x\,dF(x) + \int_{0}^{\infty} x\,dF(x) = \int_{-\infty}^{0} x\,dF(x) - \int_{0}^{\infty} x\,d\big[1 - F(x)\big]. \tag{3}$$

Using the integration by parts, we have

$$\int_{-\infty}^{0} x\,dF(x) = \Big[xF(x)\Big]_{-\infty}^{0} - \int_{-\infty}^{0} F(x)\,dx \tag{4}$$

and

$$\int_{0}^{\infty} x\,d\big[1 - F(x)\big] = \Big[x\big\{1 - F(x)\big\}\Big]_{0}^{\infty} - \int_{0}^{\infty} \big[1 - F(x)\big]\,dx. \tag{5}$$

Applying Lemma 2 to both (4) and (5), we have

$$\int_{-\infty}^{0} x\,dF(x) = -\int_{-\infty}^{0} F(x)\,dx \quad \text{and} \quad \int_{0}^{\infty} x\,d\big[1 - F(x)\big] = -\int_{0}^{\infty} \big[1 - F(x)\big]\,dx. \tag{6}$$

Substituting (6) into (3) gives

$$E(X) = \int_{0}^{\infty} \big[1 - F(x)\big]dx - \int_{-\infty}^{0} F(x)dx.$$

Since $\int_{-\infty}^{0} F(x)dx = \int_{0}^{\infty} F(-x)dx$, we have

$$E(X) = \int_{0}^{\infty} \big[1 - F(x) - F(-x)\big]dx.$$

which completes the proof. □

It should be noted that Lemma 3 is also valid for discrete random variables, but the proof is omitted.

Lemma 4. *Let $X_1, X_2, \ldots, X_n$ be a random sample with the cdf $F(x)$. Let $F_{(j)}(x)$ denote the cdf of the j-th order statistic $X_{(j)}$. Then we have*

$$\text{(i) } F_{(n)}(x) = \big[F(x)\big]^n \quad \text{and} \quad \text{(ii) } F_{(1)}(x) = 1 - \big[1 - F(x)\big]^n. \tag{7}$$

Proof. The proof is omitted. □

Theorem 5. *Let $X_1, X_2, \ldots, X_n$ be a random sample with cdf $F(x)$. Then the expectation of the range, $R = X_{(n)} - X_{(1)}$, is given by*

$$E\big(X_{(n)} - X_{(1)}\big) = \int_{-\infty}^{\infty} \Big\{1 - \big[F(x)\big]^n - \big[1 - F(x)\big]^n\Big\} dx.$$

Proof. Using Lemma 3, we have

$$E\big(X_{(n)}\big) = \int_0^{\infty} \big[1 - F_{(n)}(x) - F_{(n)}(-x)\big] dx$$

and

$$E\big(X_{(1)}\big) = \int_0^{\infty} \big[1 - F_{(1)}(x) - F_{(1)}(-x)\big] dx.$$

Applying (7) in Lemma 4 to the integral above, we obtain

$$E\big(X_{(n)}\big) = \int_0^{\infty} \Big\{1 - \big[F(x)\big]^n - \big[F(-x)\big]^n\Big\} dx,$$

and

$$E\big(X_{(1)}\big) = \int_0^{\infty} \Big\{\big[1 - F(x)\big]^n dx - 1 + \big[1 - F(-x)\big]^n\Big\} dx.$$

Thus, we have

$$\begin{aligned}
&E\big(X_{(n)} - X_{(1)}\big) \\
&= \int_0^{\infty} \Big\{1 - \big[F(x)\big]^n - \big[F(-x)\big]^n - \big[1 - F(x)\big]^n + 1 - \big[1 - F(-x)\big]^n\Big\} dx \\
&= \int_0^{\infty} \Big\{1 - \big[F(x)\big]^n - \big[1 - F(x)\big]^n\Big\} dx + \int_0^{\infty} \Big\{1 - \big[F(-x)\big]^n - \big[1 - F(-x)\big]^n\Big\} dx.
\end{aligned}$$

Using the change of the integration variable technique for the last term in the above, we have

$$\int_0^{\infty} \Big\{1 - \big[F(-x)\big]^n - \big[1 - F(-x)\big]^n\Big\} dx = \int_{-\infty}^{0} \Big\{1 - \big[F(x)\big]^n - \big[1 - F(x)\big]^n\Big\} dx.$$

It is immediate from this result that we have

$$\begin{aligned}
&E\big(X_{(n)} - X_{(1)}\big) \\
&= \int_0^\infty \Big\{1 - \big[F(x)\big]^n - \big[1 - F(x)\big]^n\Big\}dx + \int_{-\infty}^0 \Big\{1 - \big[F(x)\big]^n - \big[1 - F(x)\big]^n\Big\}dx \\
&= \int_{-\infty}^\infty \Big\{1 - \big[F(x)\big]^n - \big[1 - F(x)\big]^n\Big\}dx,
\end{aligned}$$

which completes the proof. □

It should be noted that the above lemmas and theorems are also valid for non-normal distributions. But, we use the results specifically in the case of the normal distribution. Now suppose that we have a random sample from a *standard* normal distribution, $Z_1, Z_2, \ldots, Z_n$ and that we want to calculate the expectation of the sample range. Then we have

$$E\big(Z_{(n)} - Z_{(1)}\big) = \int_{-\infty}^\infty \Big\{1 - \big[\Phi(z)\big]^n - \big[1 - \Phi(z)\big]^n\Big\}dz.$$

Note that the integrand, $1 - \big[\Phi(z)\big]^n - \big[1 - \Phi(z)\big]^n$, is an even function due to the fact that $\Phi(-z) = 1 - \Phi(z)$ which allows for the simplification of the expectation:

$$d_2(n) = E\big(Z_{(n)} - Z_{(1)}\big) = 2\int_0^\infty \Big\{1 - \big[\Phi(z)\big]^n - \big[1 - \Phi(z)\big]^n\Big\}dz.$$

Thus, the estimator $R/d_2(n) = (X_{(n)} - X_{(1)})/d_2(n)$ is unbiased for σ with $X_i \sim N(\mu, \sigma^2)$. Then it is easily seen that $R = X_{(n)} - X_{(1)} = \sigma(Z_{(n)} - Z_{(1)})$, where $Z_i \sim N(0,1)$.

Next, we consider the factor $d_3(n)$ which is defined by $\mathrm{Var}(R/d_3(n)) = \sigma^2$. Then, using $\mathrm{Var}(R) = \sigma^2\mathrm{Var}(Z_{(n)} - Z_{(1)})$, we have

$$d_3(n) = \sqrt{\mathrm{Var}(R)} = \sqrt{E(R^2) - \{E(R)\}^2} = \sqrt{E(R^2) - d_2(n)^2},$$

where $E(R^2)$ can be obtained by (2).

The value of $d_3(n)$ is involved with the double integration as shown in (2). We calculated the value of $d_3(n)$ using the numerical integration. This double integration is accurate for small values of n (say, $n < 100$). As seen in Figure 1, the numerical calculation of $d_3(n)$ (denoted by red line) is not reliable especially for large values of n. We double-checked these values by comparing with those obtained by the Monte Carlo simulation. To obtain more reliable and accurate calculation of $d_3(n)$ especially for large values of n, we calculated the value of $d_3(n)$ by Monte Carlo simulation, we generated a sample of size n from $N(0,1)$ and calculated the range. We iterated this simulation one hundred million times ($I = 10^8$) and then calculated the empirical variance of this range (denoted by a circle in Figure 1).

We also obtained the approximation of $d_3(n)$ by using the least squares method with the empirical variances of the ranges, which is given by

$$d_3(n) \approx \exp\Big(0.73784298 + 0.06390565 \cdot \ln(n) - 0.71491753 \cdot \sqrt{\ln(n)}\,\Big). \tag{8}$$

Figure 1 clearly shows that this approximation is very close to the simulated value of $d_3(n)$ for large values of n. It is worth mentioning that The `factors.cc` function calculates $d_3(n)$ using the numerical integration for $n \leq 100$ and using the approximation in (8) for $n > 100$.

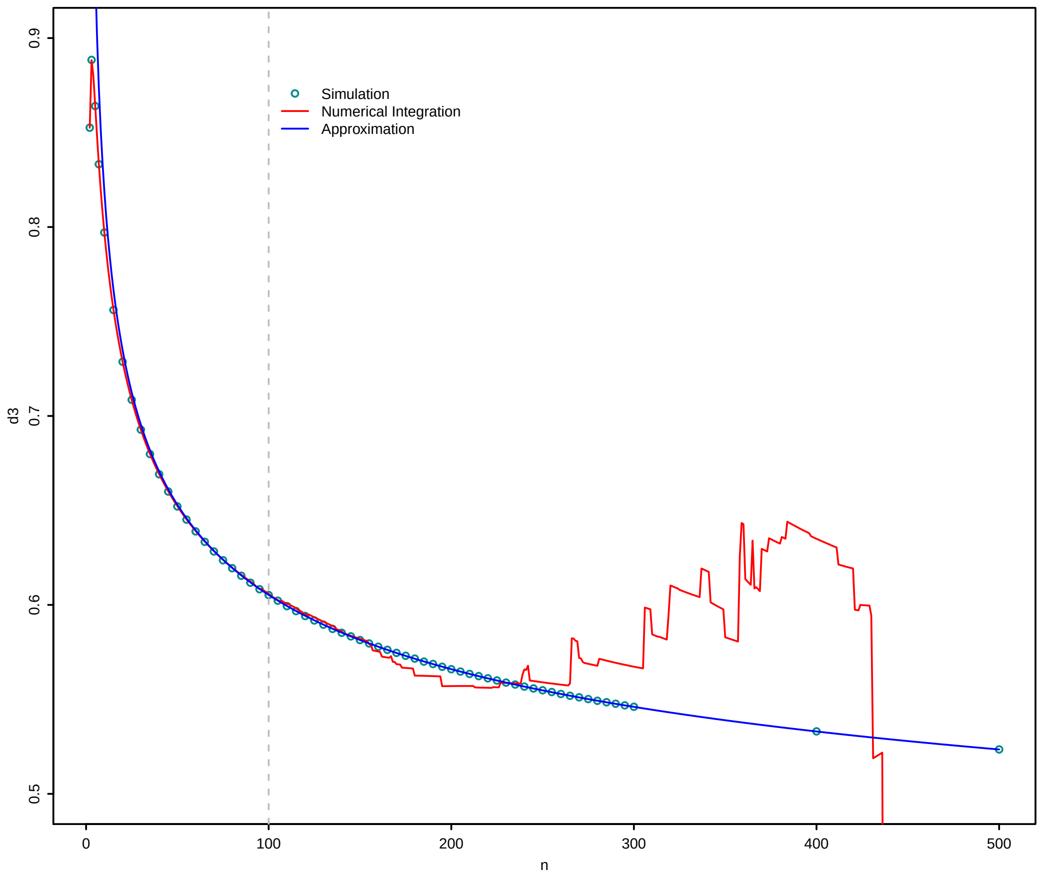

Figure 1: The plot of $d_3(n)$ versus n.

References

[1] ASTM Committee E-11. *Manual on Presentation of Data and Control Chart Analysis (STP 15-D)*. American Society for Testing and Materials, Philadelphia, PA, 4th edition, 1976.

[2] ASTM Committee E-11. *ASTM Manual on Quality Control of Materials (STP 15-C)*. American Society for Testing and Materials, Philadelphia, PA, 1951.

[3] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, second edition, 2002.

[4] H. Leon Harter. Tables of range and Studentized range. *The Annals of Mathematical Statistics*, 31(4):1122–1147, 1960.